# Public Affairs 60:
# Using Data to Learn about Society:
# Introduction to Empirical Research and Statistics

University of California, Los Angeles
Spring, 2021

## Course Information

Instructor: Dr. Marika Csapo [CHOP-oh]
Email: mcsapo@ucla.edu
Pronouns: *she/her/hers*
Class Virtual Meetings: M/W, 2:00 - 3:15 pm PST via Zoom
Virtual Office Hours: F, 1:00 - 2:15 pm (or by appt) via Zoom

Teaching Assistants:

| Name | Email | Section | | |
|---|---|---|---|---|
| Hao Ding | haoding@ucla.edu | 1B | Th | 9 - 10:50a |
| | | 1D | F | 9 - 10:50a |
| Jae Hyeon Park | jaehpark@ucla.edu | 1A | W | 4 - 5:50p |
| | | 1C | Th | 1 - 2:50p |

## Course Description

This is a class in data literacy and data analysis for students interested in social problems.

With the digitalization of a great portion of our lives in the information era, data on human behavior are being documented in ways that were unimaginable to our parents. Technological advancements in computational power have also given society the ability to process very large amounts of data toward various ends—not always in a particularly socially responsible fashion. There are positives and negatives to this trend, but whether we like it or not, social data collection and use has become an important part of our society and it looks to remain this way.

Accordingly, "data-driven" solutions to problems and "evidence-based" practices are increasingly in demand in both the private and public sector (and have been in the social sciences for quite some time). Policy proposals frequently include preliminary evidence suggesting plausible policy success. Data-collection and policy evaluation are often built into policy implementation procedures. Historical data are even used now to simulate and predict social, environmental and medical outcomes tens of years into the future—predictions that are used to inform policy decisions. Since we cannot escape data, we should learn how to use them and to use them responsibly. The tools you will acquire in this class will be very broadly useful to almost any type of inquiry.

The course will introduce students to a free open-source statistical software called "R." Students will use R to learn how to execute abstract statistical ideas in a practical way by applying them to real datasets. In the course, students will also engage in exploratory and hypothesis-oriented analysis on a topic of their choosing. The class is largely assignment and activity based because the best way to learn data analysis is to do it.

Since this is a class in data analysis, there will be some math involved. However, the prior math training you will need to feel comfortable in this class is minimal (arithmetic and basic algebra) and you will have the help of "R" to do a lot of the math for you. The course does not require any prior coding training either—just basic familiarity with computers. For those who fear math or coding, I ask you to stay open-minded—for most people math is not fun until they need it to solve a real problem they care about. This class will help you put math to good use.

## Learning Objectives

The goal of this class is to facilitate introductory-level knowledge and practice of statistical analysis, interpretation, and graphics. It also aims to encourage careful and critical consumption of quantitative information.

In this course, you will learn to:

- summarize data in a way that distills useful information about the world.
- use a flexible and broadly applicable statistical software to collect, merge, clean, and manage data.
- connect research and policy questions to measurable outcomes.
- evaluate the value and limitations of quantitative claims made by others about the world.
- make analyses replicable to others and to understand the value of transparency in research.
- differentiate between data-mining and exploratory analyses that produce hypotheses to be cross-validated.

## How You Will Be Evaluated

The assignment and evaluation structure of this class is based on the philosophy that the best way to get good at data analysis is to jump right in and do it. The marginal returns to reading about how to do statistics or watching someone else compute statistics for an hour are incredibly small compared to the returns to spending that hour doing and interpreting data analysis yourself. Therefore, homework will be the most important study tool for this class. In putting more weight on homework and projects and less weight on exams, I hope to discourage binge consumption of the material, which will not serve you well in the long-run. What you get out of this class will be proportional to how much effort you put into the homework and activities. Your focus for the homework should not just be on documenting your results, but also on carefully articulating and interpreting those results.

To do well in the class and get the most out of the material, you need to attend lecture (live, if possible, or watch the recording, if not) and section regularly. The homework assignments will help give you a sense for how prepared you are for the exam and project. Please attend office hours and sections to discuss conceptual issues, including where you may have missed points on assignments, and make sure to post questions about R or logistics on the course Slack forum where everyone can see the response (or provide their own).

I post lecture slides and recordings on the website for your convenience. These are meant to complement, not replace, lecture. I will aim to post these the same day as the lecture.

Unless otherwise stated, assignments are due by 11:59 pm the night of the due date. Most of the assignments are due on Sundays, but your reflections on data and analysis are due on Fridays. The due dates for all assignments are listed in the course lecture schedule below. Completed assignments should be uploaded to CCLE through the appropriate link by the due date.

**Course Grading Breakdown**

| Course Component | Contribution to Grade |
|---|---|
| Labs (2) | 10% |
| Homework Assignments (4) | 30% |
| Research Activities (4) | 35% |
| Exam (1) | 15% |
| Participation (4) | 10% |

Brief descriptions of the graded course components follow.

**Labs** - 10% There are two labs, which focus on coding training early in the quarter and can be completed in section during the week they are assigned.

- Lab 1 - "Getting Comfortable with R" - 5%
- Lab 2 - "Using R to Clean and Summarize Data" - 5%

**Homework Assignments** - 30% There are four homework assignments. These provide opportunities to apply lecture materials and statistical techniques to real data and interpret the results. You may get help from TAs in sections, from myself or TAs in office hours, and you may discuss the problems, approaches, and interpretations with friends. However, the code and the write-up that you turn in must be uniquely your own and written in your own voice and diction (no identically-worded submissions). Practice and feedback on these assignments will help you in your research activities.

- Homework 1 - "Getting Comfortable with Summaries and Graphics in R" - 5%
- Homework 2 - "Bivariate Regression Interpretation" - 10%
- Homework 3 - "Regression Assumptions and Multivariate Regression" - 10%
- Homework 4 - "Question-Driven Data Analysis Practice" - 5%

**Research Activities** - 35% There are four research activities. These are geared toward putting practice into action on an issue that is more personal to you. These will be very individualized analyses and not everyone will use the same dataset. Consultation with TAs, myself, classmates, your mom and your bestie are allowed and encouraged. However, submitted write-ups must be uniquely your own (written in your own voice and diction).

- Research Activity 1 - "Constructing and Documenting Your Own Dataset" - 10%
- Research Activity 2 - "Preliminary Research Proposal" - 10%
- Reseach Activity 3 - "Results Write-up with Reproducible Code" - 10%

- Research Activity 4 - "Peer Review and Replication" - 5%

**Exam** - 15% There is one (final, cumulative) exam. The exam will be made up of a combination of multiple choice and short answer questions focused on your conceptual understanding and interpretation of statistical output. You will not be asked to code anything for the exam. The exam will be open-note (lecture slides, recordings, notes, reading materials are fine to use). However, you may NOT collaborate on the exam or consult anyone on exam materials (save for clarifying questions for TAs or myself).

**Participation** - 10% There are four grade components based on participation of various kinds (all are credit/no credit).

- Participation 1 - "Student Survey" - 1%
- Participation 2 - "Data Share and Reflections" - 3%
- Participation 3 - "Discussion Forum and/or Section Participation" - 5%
- Participation 4 - "Anonymous Course Evaluation Participation" - 1%

**Grade Scale** At the end of the quarter the weighted-average will be computed over the components according to the weighting scheme in described in the Table above. This weighted average score for the quarter will be used to determine the final letter grade for the course using the mapping in Table 1 below.

Table 1: Grade Scale

| **Letter Grade** | **Range** |
|---|---|
| A+ | $97 \rightarrow 100$ |
| A | $93 \rightarrow 97$ |
| A- | $90 \rightarrow 93$ |
| B+ | $87 \rightarrow 90$ |
| B | $83 \rightarrow 87$ |
| B- | $80 \rightarrow 83$ |
| C+ | $77 \rightarrow 80$ |
| C | $73 \rightarrow 77$ |
| C- | $70 \rightarrow 73$ |
| D | $60 \rightarrow 70$ |
| F | $0 \rightarrow 60$ |

## Course Materials

**Course Website**: You can find our CCLE course website at `https://ccle.ucla.edu/course/view/21S-PUBAFF60-1`.

**Reading Materials**: There are no assigned texts for this class. Every Thursday I will post reading for the following week. I will include an optional textbook-like resource on that week's topic for those who feel they could use some more formal reinforcement of lecture material. I will also post current affairs articles and suggested discussion topics (please feel free to reflect on discussion topics in the Slack forum to pick up participation credit). You will be responsible for the current affairs and example articles which are fair game for exam materials. If you feel compelled to purchase a formal text, consider Moore and McCabe's *Introduction to the Practice of Statistics*, any edition. If you find yourself struggling with R and would like a handy reference, check out Grolemund and Wickham's *R for Data Science*, which is available for free online.

**Discussion Forum**: I will set up a discussion forum on Slack, which you will be invited to join. All logistics questions and conceptual and coding questions for Labs and Homework should be posted there so everyone can benefit from the responses. Every week a couple of students from each section will post a news, research, blog or other article type and an evaluation of the analysis. I will choose one or two of these to highlight on occasion. You will not be tested on the contents of these articles, but you are welcome to pick up participation points by posting comments to Slack to engage the reflections of your peers when they share an article.

**Hardware and Software**: You need access to a computer and wifi to participate in lectures and sections, download course materials, and upload assignments to CCLE. The software for class include R and RStudio. Information on download can be found at
`http://cran.r-project.org/` for R and at
`http://www.rstudio.com/products/rstudio/download/` for RStudio. More detailed information about downloading the software appear in the instructions to Lab 1.

**Datasets**: In this course you will be doing an exploratory data project on a topic of your choosing. I will be linking a document with social science data resources for various topics to give you dataset options to choose from. You are welcome to find and include data from sources I have not listed (but if so, please consult with instructor or TA to get approval first).

## Wellness Resources

This is an unusual quarter with a lot of new logistical, financial, temporal, and psychological demands placed on students (as well as faculty). I will do my best to accommate individual demands.

**COVID-19**:

- If you experience COVID-like **symptoms**, please call the UCLA Ashe Center COVID Hotline at (310) 206-6217 to get more information and resources.

- You can find up-to-date information on **campus services, resources, and policies** during COVID-19 at `https://covid-19.ucla.edu/information-for-students/`.
- Information about COVID-19 **testing** through Los Angeles County can be found at `https://covid19.lacounty.gov/testing/`.
- LA County provides information about **food assistance** for those facing food insecurity during COVID-19 here: `https://covid19.lacounty.gov/food/` and other resources for the vulnerable populations including some pertaining to **eviction prevention** here: `http://www.publichealth.lacounty.gov/media/Coronavirus/resources.htm`.

**Other Wellness Resources**:

- One number to keep in your phone contacts is UCLA **Counseling and Psychological Services** (CAPS) 24-hour number for mental health support (as well as scheduling during daytime hours): 310-825-0768. CAPS offers individual services as well as many support and self-care groups for students. Find out more about the services they offer at `https://www.counseling.ucla.edu/`.
- Useful resources for students in crisis are available at `https://www.studentincrisis.ucla.edu/`.
- Resources for students to promote online learning can be found at `https://www.teaching.ucla.edu/resources/student-remote-learning`.
- UCLA Office of Diversity and Inclusion resources for **race-related trauma**: `https://equity.ucla.edu/know/resources-for-racial-trauma/`. CAPS also offers individual and group counseling specifically oriented toward racial trauma (see above for phone and website). Finally, Community Care for Black Bruins is available through UCLA's Rise Center at `https://risecenter.ucla.edu/virtual-library/healing-support-for-black-bruins`.
- **Confidential legal counseling** is available for UCLA students through `https://www.studentlegal.ucla.edu/` including eviction prevention advice and consultations regarding the rights of students who are either documented or undocumented immigrants.
- Accommodations requests for **students with disabilities** are welcome and may be made through the Center for Accessible Education (CAE). Please see the CAE website, `https://www.cae.ucla.edu/`, for more information about how to request accommodations.
- UCLA students have access to online **guided meditations, peer support, and mindfulness resources** through UCLA's RISE program. Go to `https://risecenter.ucla.edu/visit-us` for more information.

## Course Policies

**Collaboration**: Students are welcome and even encouraged to discuss homework problems and R techniques outside of class. However, each student MUST do their own write up,

using their own examples and their own interpretation of the results in their own voice and diction. Turning in the same write up as another student (all or in part) on any assignment for this class will be considered academic dishonesty. Consequences depend on the severity of the infraction but range from a score of zero on the assignment to a university investigation and even potential expulsion for extreme cases. Collaboration on exams is not allowed in any form and will also be considered academic dishonesty.

**Inclusion Statement**: One of the tremendous benefits of a UCLA education is the diversity of experiences, identities, and perspectives represented in our classrooms. All of these experiences are institutional assets as they are indispensable to the goal of challenging students to think critically. Exposure to new ideas and experiences that force us to thoroughly explore and clearly articulate own ideas promotes conscientious belief-formation and positive social change. Therefore, openness to dialogue and respect for the asset which is our diversity is important and I will do all that I can to create a space where all feel safe to participate. Toward that end, discussions will not be laissez-faire. One's right to free speech will be limited in that it may not be used to decrease anyone else's safety in participating, thereby diminishing their access to free speech. Hate speech will not be tolerated in the classroom or the University and respect for identity autonomy is non-negotiable.

**Grade Disputes**: Grade disputes must be made in writing, should be made only after requesting feedback from the TA, and must make clear which question or portion of the assignment you believe deserved a better grade and, most importantly, why. Regrade requests will be granted under these circumstances, but assignments will be re-graded in their entireity. This could lead to an increase or decrease in total points allocated so please evaluate whether a regrade is likely to produce an improvement before requesting it.

## Lecture and Section Schedule and Due Dates

## Week 1

**Lecture: Monday, March 29.**

- Syllabus review.
- Why data analysis?
- Course outline.
- R and RStudio.

**Lecture: Wednesday, March 31.**

- Random Variables - Univariate.
- Conceptualizing Dataframes.
- Types of Data.

**Section.**

- R and RStudio download and troubleshoot.
- Lab 1 workshop.
- Data share Sign-up

**Due Dates.**

- Lab 1 (by 11:59 pm Sunday, April 4).
- Student Survey (by 11:59 pm Sunday, April 4).

## Week 2

**Lecture: Monday, April 5.**

- Distributions: Nominal and Ordinal Data.
- Distributions: Continuous Data.
- Mean and Conditional Mean

**Lecture: Wednesday, April 7.**

- Measures of Central Tendency and Outliers.
- Percentiles.
- Other single-number summaries.
- Boxplots and Multi-number summaries.
- Measures of Spread for a Single Random Variable.

**Section.**

- Lab 2 workshop.

**Due Dates.**

- "Data Share" Group 1 (by 11:59 pm Friday, April 9).
- Lab 2 (by 11:59 pm Sunday, April 11).

## Week 3

**Lecture: Monday, April 12.**

- Conceptualizing the mean as satisfying min SSR.

**Lecture: Wednesday, April 14.**

- Summaries of the relationship between two random variables.
- Covariance, Correlation.

**Section.**

- Homework 1 workshop.

**Due Dates.**

- "Data Share" Group 2 (by 11:59 pm Friday, April 16).
- Homework 1 (by 11:59 pm Sunday, April 18).

## Week 4

**Lecture: Monday, April 19.**

- Partitioning variance.
- ANOVA.

**Lecture: Wednesday, April 21.**

- Introduction to Regression Analysis.
- Bivariate OLS as a conditional mean minimizing SSR.
- Partitioning into predicted values and residuals.

**Section.**

- Homework 2 workshop.

**Due Dates.**

- "Data Share" Group 3 (by 11:59 pm Friday, April 23).
- Homework 2 (by 11: 59 pm Sunday, April 25).

## Week 5

**Lecture: Monday, April 26.**

- Comparing OLS to correlation coefficient.
- Comparing OLS to ANOVA.
- Diagnosing Problems with OLS.
- Leverage.
- Influence.

**Lecture: Wednesday, April 28.**

- Research Questions and Operationalization.
- Non-linear variable transformations.

**Section.**

- Research Activity 1 workshop.

**Due Dates.**

- "Data Share" Group 4 (by 11:59 pm Friday, April 30).
- Research Activity 1 (by 11:59 pm Sunday, May 2).

## Week 6

**Lecture: Monday, May 3.**

- Introduction to Multivariate Regression.
- R-Squared

**Lecture: Wednesday, May 5.**

- Compare Multivariate OLS to stepwise regression.
- Compare Multivariate OLS to partial regression.

**Section.**

- Homework 3 workshop.

**Due Dates.**

- "Data Share" Group 5 (by 11:59 pm Friday, May 7).
- Homework 3 (by 11:59 pm Sunday, May 9).

## Week 7

**Lecture: Monday, May 10.**

- Dummy Variables.
- Thinking about time.

**Lecture: Wednesday, May 12.**

- Generalized Linear Models.
- Cross-validation.

**Section.**

- Research Activity 2 workshop.

**Due Dates.**

- "Data Share" Group 6 (by 11:59 pm Friday, May 14).
- Research Activity 2 (by 11:59 pm Sunday, May 16).

## Week 8

**Lecture: Monday, May 17.**

- Introducing inference.
- Monte Carlo methods and Computational Social Science.
- Populations and Samples.
- Sampling Distributions and the Central Limit Theorem.

**Lecture: Wednesday, May 19.**

- Standard errors.
- z- and t- values.
- Confidence intervals, Margin of Error.
- Hypothesis testing and p-values.
- Type I and Type II Errors.

**Section.**

- Homework 4 workshop.

**Due Dates.**

- "Data Share" Group 7 (by 11:59 pm Friday, May 21).
- Homework 4 (by Sunday, May 23).

## Week 9

**Lecture: Monday, May 24.**

- Returning to Research Design.
- Observational data, surveys, missingness.
- Causality v. Correlation
- Randomization, experimentation, field v lab experiments.
- Design for Policy Evaluation.

**Lecture: Wednesday, May 26.**

- Loose ends.

**Section.**

- Research Activity 3 workshop.

**Due Dates.**

- "Data Share" Group 8 (by 11:59 pm Friday, May 28).
- Research Activity 3 (by 11:59 pm Sunday, May 30).

## Week 10

**Lecture: Monday, May 31.**

- Memorial Day. NO CLASS!

**Lecture: Wednesday, June 2.**

- Summation and Exam Review.

**Section.**

- Exam Review and Research Activity 4 workshop.

**Due Dates.**

- "Data Share" Group 9 (by 11:59 pm Friday, June 4).
- Research Activity 4 (by 11:59 pm Sunday, June 6).

## Finals Week

**Due Dates.**

- Final Exam window opens for 24-hours on Tuesday, June 8 at 3 pm (closes Wednesday at 3 pm).